Data Mining Techniques on Traffic Violations

Santosh Thapa, Advisor: Prof. Jeongkyu Lee
Department of Electrical and Computer Engineering, University of Bridgeport, CT.

Abstract

This paper describes use of Data mining techniques used to model traffic accidents detection. It is done by determing the blackspots by using Association Rule Mining and Clusterization algorithm. It helps to ascertain the traffic violation patterns and blackspot of traffic violations. We looked into K-means clustering with some enhancements to aid in the process of identification of patterns and blackspots. We applied these techniques to real traffic data extracted from the Montgomery County of Maryland and validated our results. We also developed a prioritized scheme for attributes here to deal with the limitations of various out of the box clustering methods and ways. This easy to implement data mining framework works with the geo-spatial plot of blackspots and helps to improve the road accidents zones.

Introduction

With the increasing number of vehicles, traffic violations occur frequently and hence an accident. There are some situations or cases where the traffic violations gets eventually transformed into accidents occuring at the same place or location more than once which could be at a sharp drop or corner. Road traffic accident hazardous locations are called black spots, i.e. the road sections or intersections where traffic accidents are high. Locating these very prone hazardous spot is the first step to improve the road safety state which is indeed a very important state. Therefore, identification of traffic black spot, analysis of the cause of traffic black spot and then reliable operations to improve traffic safety state in short duration of time. But, traffic violations are mainly the causes which render traffic accidents. If traffic violations can be controlled efficiently, traffic accidents can be reduced remarkably. Taking the reference of the traffic violations which caused accidents, traffic violation black spots can be obtained. It can reduce traffic violations efficiently through improving the traffic condition of traffic violation black spots.

Many data mining methods can be used to forecast traffic violation, such as linear regression method, exponential smoothing method, Kalman filtering method, BP neural network, Support Vector Machines(SVM) and so on. SVM is popular network used for classification and regression. In the SVM method, it is a burdensome work to fix the kernel parameters. Bayesian interference can improve the process markebly. We will utilize the nonlinear regression SVM whose kernel parameters are fixed with Bayesian interference method to forecast traffic violations. Then, proposed identification method is used to identify traffic violation black spots.

Goal/Objectives

There are several public transportations like bus, shuttles, cab and so on as well which are all summing up to make a huge number of vehicles on the streets and highway. But, even when the road conditions are perfect, a traffic violations could turn into a hazardous accident.

Road traffic safety deals with a complexity of problems as there are some factors that contribute to its disturbance: blind curve, sharp corner, width of roads, driver's behaviour in traffic, wear and tear of infrastructure, climate scenarios, light conditions and intense traffic. Due to continuous increase of traffic, the amount of accidents rises notably while road traffic congestion safety becomes more hard to maintain.
But, out of all, there are most accident prone locations also called black spots where traffic violations turn out to be road accidents. Locating this particular accident hazardous spot is the first step to improve the road safety state. In addition to that, we can also predict a particular age-group of individuals who are mostly making these violations.

And, a specific ethnic group or age-group who mostly cause the traffic violation is also determined.

Data Set

The dataset contains traffic violation information from all electronic traffic violations issued in Montgomery County, Maryland and was extracted from Montgomery County of Maryland, data.montgomerycountymd.gov. It includes all traffic violations since January 1, 2012, so, this analysis is based on 49 months of records.
There are 35 attributes and total 819565 instances on this dataset.

The main attributes that'll be used for data mining are Location, Geo-location, Accident, Alcohol, Work Zone, Year, Violation Type, Charge, Gender, Race and so on. The attribute, geo-location and location will be used to visualize in a geo-spatial data visualization.

Data Mining Algorithms

The procedure of Traffic violation black spot forecasting can be summarized as follows. Firstly, history traffic violation records should be collected. Select the records of a road section or intersection within the analysis time period. The data are further pre-processed using specific filters in Weka. Hidden best rules are determined using Association Rules mining. Then, by using the K-means Algorithm a specific clusterized data is obtained. After training, SVM is used to forecast the future traffic violation number. Every road section and intersection is processed with the procedure.

Support Vector Machines(SVM) are managed learning prototype with related learning algorithms that inspects data used for sorting and regression study. Given, a set of training samples, each marked for relating to one of two groups, an SVM training algorithm forms a replica that assigns new samples into one group or the other, making it a non-probabilistic binary linear distributer

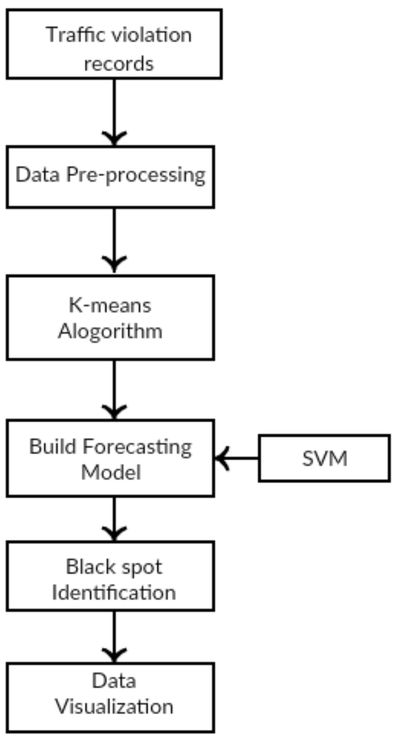


Figure : Algorithm

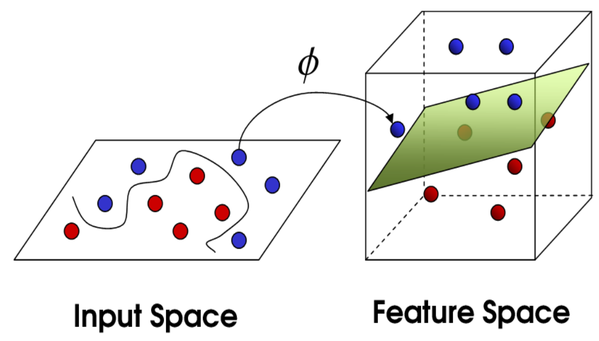
Evaluation

The main objective of the project is to determine the black spot of traffic violations. It can be determined by using k-means algorithm. The objective is to allow a cluster to each and every data point. K-means is a clustering method that desires to discover the positions μi,i=1...k of the clusters that minimize the square of the length from the data points to the cluster. K-means clustering solves-

$$\arg min_s \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2$$

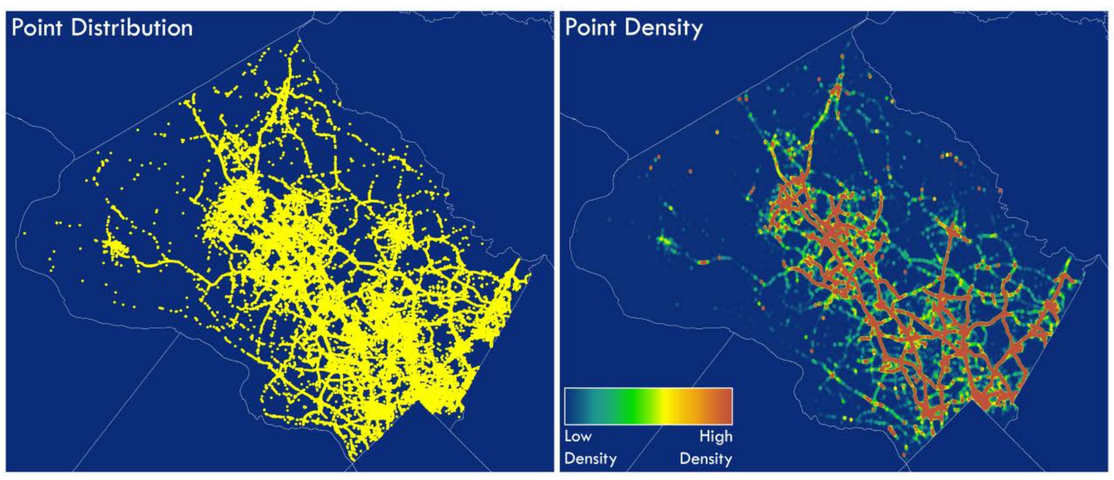
where, S_i is the set of points that is located to cluster i. The K-means clustering uses the Euclidean distance d(x,μi)=‖x−μi‖2. This problem is not insignificant (in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution.

Support Vector Machines(SVM) is used to determine the future traffic violations. Nonlinear regression SVM first maps the input points into a high-dimensional feature space with a nonlinear mapping function ϕ and then conducted via linear regression in the high-dimensional feature space. The aligned regression in high-dimension feature space corresponds to the non-aligned regression in low-dimensional input space.



Data Visualization

Data visualization is the main theme on the data mining. There are six representations performed which were:-
1. Geo-location representation showing point distribution and density of traffic violations incurred.
2. A graphical portrayal comparing how traffic violations are distributed among races (white, black, hispanic, asian & so on) and sex(male & female).
3. A simple presentation of the ratio of citations per warning for each demographic. When a violation occurs, a person is either issued a citation, a warning, or a safety equipment repair order.
4. A deployment of a violations by age or race by car model year.
5. A time series of when each race commits violation throughout the day.
6. Considering all the speeding violations, a depiction of these records in a bar-graph. It will show, for each speed limit, how many miles per hour over the limit the person was driving.



Conclusions

The prediction of a traffic violation black spot is a very important issue in traffic safety. SVM is used to find the forecasting model. Similary, Association rules mining and K-means clusterization data mining techniques are used to find out the hidden relational rules and clusterized data respectively. This project will indeed make people to be extra careful before entering into the accident prone zones and drive steadily and carefully. This concept can be regulated into the GPS module of the vehicles as a geo-location representation of blackspots.

Future Work

Potential future enhancements to this project include:
- Installation of the blackspot program on the vehicles of GPS system.
- Regulate on the public transport services

References

1. Sheng Yugang, Xu Weijuan, "Optimization Model of Traffic Accident Black-Spots Formation Mechanism," Road traffic and safety, vol.2,pp. 80-82, 2010
2. Haifeng Jiang, Liande Zhong, Changcheng Li, Han Feng, "Research on identification method for road accident blackspots with ordinal clustering method," Remote Sensing, Transportation Engineering, vol18, pp. 2401-2404, 2011.
3. Xing Dawei Li Xiansheng, "Identification of Speedway Accident Black Spots based on the Quality Control method," International Conference, pp 541-544, 2015.
4. Jihua Ye, Yanhui Zhou, Ming Li, Chunlan Wnag, "Research and implement of traffic accident analysis system based on accident black spot," Road traffic and safety, vol.7, pp1805-1809, 2010.